PCT/GB 00/01715

The Patent Office

PATENTS • DESIGNS • COPYRIGHT • TRADE MARKS

10/030524

INVESTOR IN PEOPLE

GB00/01715

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

REC'D 20 JUN 2000

WIPO PCT

4

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed

Dated 12 JUN 2000

THIS PAGE BLANK (USPTO)

Patents Form 1/77

Patents Act 1977
(Rule 16)

07MAY99 E445132-5 D01559
P01/7700 0.00 - 9910448.1

The Patent Office

THE PATENT OFFICE
K
- 7 MAY 1999
RECEIVED BY POST

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

---

1. Your reference — BKCD/IRJ/ENS.7

2. Patent application number *(The Patent Office will fill in this part)* — **9910448.1**

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*

   Ensigma Limited,
   Turning House,
   Station Road,
   Cheptstow. NP6 6PB

   Patents ADP number *(if you know it)* — 05723168002.

   If the applicant is a corporate body, give the country/state of its incorporation

4. Title of the invention — Cancellation of Non-Stationary Interfering Signals for Speech Recognition

5. Name of your agent *(if you have one)*

   "Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

   Wynne-Jones, Laine & James
   22 Rodney Road,
   Cheltenham,
   Glos. GL50 1JJ

   Patents ADP number *(if you know it)* — 1792001 ✓

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

Country	Priority application number *(if you know it)*	Date of filing *(day / month / year)*
n/a		

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application	Date of filing *(day / month / year)*
n/a	

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*
   a) any applicant named in part 3 is not an inventor, or
   b) there is an inventor who is not named as an applicant, or
   c) any named applicant is a corporate body.
   See note (d))

   YES

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form	–
Description	29
Claim(s)	7
Abstract	1
Drawing(s)	5 + 5

10. If you are also filing any of the following, state how many against each item.

Priority documents	–
Translations of priority documents	–
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	–
Request for preliminary examination and search *(Patents Form 9/77)*	–
Request for substantive examination *(Patents Form 10/77)*	–
Any other documents *(please specify)*	–

11. I/We request the grant of a patent on the basis of this application.

Signature: Wynne-Jones Lainé & James

Date: 6th May, 1999

12. Name and daytime telephone number of person to contact in the United Kingdom

B K C DUNLOP 01242 515807

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) *If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.*

b) *Write your answers in capital letters using black ink or you may type them.*

c) *If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.*

d) *If you have answered 'Yes' Patents Form 7/77 will need to be filed.*

e) *Once you have filled in the form you must remember to sign and date it.*

f) *For details of the fee and ways to pay please contact the Patent Office.*

Cancellation of Non-Stationary Interfering Signals for Speech Recognition

This invention relates to apparatus and method for cancellation of non-stationary interfering signals. In particular, the invention relates to cancellation of such signals for the purpose of recovering a wanted speech signal for use by a speech recognition application. The invention is especially suitable for use in an automobile where in-car devices produce interfering signals during the speech recognition process.

A problem associated with speech recognition is that of maintaining performance in the presence of interfering signals so that the speech recognition process continues to function satisfactorily even in the presence of background noise. Known systems have been directed towards mitigating effects of quasi-stationary noise such as telephone channel noise or car noise. Proposed solutions to quasi-stationary noise interference include spectral subtraction, Weiner filtering and parallel model combination, each of which work in the spectral domain.

There are, however, other sources of interference in acoustic environments which may degenerate performance of speech recognition applications. In the example of an automobile environment, in addition to engine noise, another source of potentially interfering non-stationary acoustic signals includes sound generated by electronic devices operating in the car. Examples of such devices include in-

car entertainment accessories such as radios, compact disc players and tape players and also other types of devices which may emit sonic signals, e.g. telephone ringing or navigation system warning tones. In this specification, electronic devices capable of emitting acoustic signals and operating in a vehicle are generically referred to as "Electronic in-car Acoustic Devices (ECAD).

Sound generated by ECAD could be present when a user wishes to control a device using a voice command. For example, a radio may be playing in a car when the user wants to use voice control of a navigation system or the radio itself. In this case, the original interfering signal produced by the radio is assumed to be known and accessible but has passed through an unknown acoustic path between the radio's loudspeakers and the speech recognition system's microphone. The acoustic path may be determined by the position of the loudspeakers and the microphone inside the car as well as other factors, such as the number of passengers and the presence of luggage inside the car.

Known systems which attempt to overcome the problem of non-stationary interferers have been based on time domain adaptive filters. However, although adaptive filtering may produce satisfactory results, this approach suffers from a number of disadvantages. Such disadvantages include high computational requirements and slow convergence of adaptive filtering algorithms. Simple forms of adaptive filtering may require order 3N computations per sample. Such high computational requirements can mean that complex hardware

may be required in order to perform the necessary filtering, thereby increasing costs of devices incorporating such technology to the consumer.

According to a first aspect of the present invention, there is provided apparatus for cancellation of one or more non-stationary interfering signals for speech recognition, said apparatus comprising:

means for receiving an acoustic signal;

means for generating an estimated value of a magnitude spectrum of said non-stationary interfering signals; and

means for subtracting said estimated value from said received acoustic signal to produce a representation of a wanted speech magnitude spectrum.

Preferably, said means for generating estimated value includes processing means configured to estimate a transfer function for an acoustic channel between each source of said non-stationary interfering signals and said means for receiving an acoustic signal.

Preferably, said processing means is configured to estimate transfer functions for non-stationary interfering signals produced by left and right stereo channel transmissions.

Preferably, said estimation of said transfer functions is achieved by said processing means executing an iterative algorithm on a frame-by-frame basis, the frames being constituted by successive time periods.

Preferably, said processing means is configured to estimate magnitudes of said left and right channel interfer-

ence signals,

said magnitude of left channel interference signal estimated by subtracting said right channel interference signal magnitude estimated during previous said iteration from said acoustic signal received at current said iteration; and

said magnitude of right channel interference signal is estimated by subtracting said left channel interference signal magnitude estimated during previous said iteration from said acoustic signal received at current said iteration.

Preferably, said transfer function estimate for said right stereo acoustic channel is determined by dividing said right channel interference magnitude estimate by said interfering signal transmitted from said right acoustic stereo channel; and

said transfer function estimate for said left stereo acoustic channel is determined by dividing said left channel interference magnitude estimate by said interfering signal transmitted from said left acoustic stereo channel.

Preferably, said right acoustic channel transfer function estimation is performed for a said iteration only if a ratio of total energy of said right acoustic stereo channel interfering signal over total energy of said left acoustic stereo interfering channel exceeds a predetermined threshold value; and

said left acoustic channel transfer function estimation is performed for a said iteration only if a ratio of total

energy of said left acoustic stereo channel interfering signal over total energy of said right acoustic stereo channel interfering signal exceeds a predetermined threshold value.

Preferably, said ratio and threshold comparisons are applied to individual frequency components in spectra of said signals.

Preferably, said left and right stereo acoustic channel transfer functions are multiplied by $(1-|\eta(k)|)$ where $\eta(k)$ is coherence of said left and right interfering signals at a frequency index k.

Preferably, said transfer function estimate for said right stereo acoustic channel is obtained using an expression:

$$\hat{H}_{AR}(k) = \frac{Y(k)}{R''(k)} = \frac{H_{AR}(k)\,.\,R''(k)}{R''(k)} = H_{AR}(k)$$

and said transfer functions estimate for said left stereo acoustic channel is obtained using an expression:

$$\hat{H}_{AL}(k) = \frac{Y(k)}{L''(k)} = \frac{H_{AL}(k)\,.\,L''(k)}{L''(k)} = H_{AL}(k)$$

wherein $R''(k)=H_{CR}(k).C(k)$, with $C(k)$ being a common component of said left and right stereo channel signals and $H_{CR}(k)$ is a transfer function between common said left and right stereo channel transmissions, and said right stereo channel and $L''(k)=L(k)-H_{CL}(k).C(k)$, where $H_{CL}(k)$ is a transfer function between common said left and right stereo channel

transmissions and said left stereo channel signal.

Preferably, wherein said processing means further comprises means for smoothing said estimated transfer functions in time domain.

Preferably, wherein said means for smoothing in time domain comprises a first order recursive filter.

Preferably, said processing means further comprises means for smoothing said estimated transfer functions in frequency domain.

Preferably, said means for smoothing in frequency domain comprises a Finite Impulse Response filter.

Preferably, said processing means includes means for performing a Fourier Transform.

Preferably, said non-stationary interfering signals are produced by an electronic acoustic device operating in a vehicle.

Preferably, said means for receiving an acoustic signal comprises a microphone.

According to a second aspect of the present invention there is provided a method of cancellation of one or more non-stationary interfering signals for speech recognition, said method comprising steps of:

receiving an acoustic signal;

generating an estimated value for a magnitude spectrum of said non-stationary interfering signal; and

subtracting said estimated value from said received acoustic signal to produce a representation of a wanted speech magnitude spectrum.

Preferably, said step of generating an estimated value comprises estimating a transfer function for an acoustic channel between each source of said non-stationary interfering signals and said means for receiving an acoustic signal.

Preferably, said transfer functions are estimated for non-stationary interfering signals produced by left and right stereo channel transmissions.

Preferably, said step of generating an estimated value is executed iteratively on a frame-by-frame basis.

Preferably, said step of estimating a transfer function includes:

estimating a magnitude of said left channel interference signal by subtracting said right channel interference signal magnitude estimated during previous said iteration from said acoustic signal received at current said iterations; and

estimating magnitude of said right channel interference signal by substracting said left channel interference signal magnitude estimated during previous said iteration from said acoustic signal received at current said iteration.

The method may further comprise steps of:

dividing said right channel interference magnitude estimate by said interfering signal transmitted from said right acoustic stereo channel; and

dividing said left channel interference magnitude estimated by said interfering signal transmitted from said left acoustic stereo channel.

Preferably, said step of estimating right acoustic

channel transfer function is performed for a said iteration only if a ratio of total energy of said right acoustic stereo channel interfering signal over total energy of said left acoustic stereo channel interfering signal exceeds a predetermined threshold value; and

said step of estimating left acoustic channel transfer function estimate is performed for a said iteration only if a ratio of total energy of said left acoustic stereo channel interfering signal over total energy of said right acoustic stereo channel interfering signal exceeds a predetermined threshold value.

Preferably, said ratio and threshold comparisons are applied to individual frequency components in spectra of said signals.

Preferably, said left and right stereo acoustic channel transfer functions are multiplied by $(1-|\eta(k)|)$ where $\eta(k)$ is coherence of said left and right interfering signals at a frequency index k.

Preferably, said transfer function estimate for said right stereo acoustic channel is obtained using an expression:

Preferably, this aspect may be comprising a step of smoothing said estimated transfer functions in time domain.

Preferably, this aspect may be further comprising a step of smoothing said estimated transfer functions in frequency domain.

According to a third aspect of the present invention, there is provided a speech recognition system including

apparatus according to the first aspect of the invention. According to a fourth aspect of the present invention, there is provided an electronic acoustic device including apparatus according to the first aspect of the invention.

The present invention is a frequency domain (rather than time domain as used in known systems) technique solution which is preferably based on channel identification followed by spectral subtraction. Embodiments of the present application's system can substantially improve performance of a speech recognition system when non-stationary interferers are present whilst having a an advantage of lower computational requirement than known systems.

Embodiments of the present application's system provides levels of non-stationary interferer cancellation sufficient to substantially improve the performance of a speech recognition system, typically about 10 decibels of cancellation is possible in the case where loud background music is being output by ECAD. Such levels of cancellation may not be satisfactory to a human listener, however, for the purposes of speech recognition applications, such levels of cancellation will substantially improve the system's performance. A human listener is sensitive to levels of interference 40 decibels below the level of wanted signal, whilst known speech recognition systems can operate well with a 15 decibel signal-to-noise ratio.

The interfering signal output by an ECAD such as a radio may be a mono or stereo transmission, typically being output from two loudspeakers located at separate locations

within an automobile. For the purposes of the description, it is generally assumed that a phase of the interferer signal is not required at the speech recognition system, as recognition feature sets such as cepstra do not normally contain phase information.

The invention may be performed in various ways and, by way of example only, a specific embodiment thereof will now be described, reference being made to the accompanying drawings, in which:

Figure 1 illustrates schematically an example of an automobile environment having ECAD where a speech recognition system is used to control an in-car device;

Figure 2 illustrates a flow diagram representing steps which may be used to estimate transfer functions representing a model of an in-car acoustic channel;

Figure 3 illustrates schematically components which may be used to implement a refinement of the algorithm in Figure 2;

Figure 4 illustrates a block diagram representing a specific embodiment of the present invention; and

Figures 5 to 8 illustrate examples of microphone signals obtained during experimental use of the present invention.

Figure 1 illustrates schematically a simple situation in which stereo ECAD signals are transmitted from separate loudspeakers. Left stereo signal $L(j\omega)$ is transmitted from left loudspeaker 101 and right stereo signal $R(j\omega)$ is transmitted from right stereo speaker 102.

Loudspeakers 101 and 102 are typically located in panelling on driver and passenger's doors. Further loudspeakers may also be fitted in the vehicle, for example they may be located in a boot compartment at the rear of the car. It will be appreciated by those skilled in the art that the specific embodiment described herein intended for use with two loudspeakers could be modified to function with different numbers of loudspeakers, which may or may not be configured to generate signals which correlate with signals being output from other loudspeakers present in the car.

Figure 1 also includes a microphone 103 which is preferably connected to an in-car electronic device such as the radio for the purpose of receiving acoustic signals which may be used by a speech recognition system for controlling the device.

A user's voice command which may be processed by the speech recognition system in order to control the electronic device is represented by wanted speech signal $S(j\omega)$104.

A spectrum of the acoustic signal received at the microphone, denoted by $Y(\omega)$, comprises components including a combination of the wanted speech $S(j\omega)$ and the signals produced by the loudspeaker having passed through an acoustic channel defined by the in-car environment.

Perfect cancellation of the unwanted ECAD stereo signals $L(j\omega)$ and $R(j\omega)$ could in principle be achieved given knowledge of acoustic transfer functions $H_{AR}(j\omega)$ for the acoustic path between the right loudspeaker 102 and microphone 103 and acoustic transfer function $H_{AL}(j\omega)$ for the

acoustic path between the left loudspeaker 101 and microphone 103. If the transfer functions $H_{AL}(j\omega)$ and $H_{AR}(j\omega)$ were known, it would be possible to retrieve a signal corresponding to the wanted speech command spoken by the user by subtracting the left stereo source signal $L(j\omega)$ transferred by $H_{AL}(j\omega)$ and the right source signal $R(j\omega)$ transferred by $H_{AR}(j\omega)$ from the signal $Y(j\omega)$ received at microphone mono 103. However, in practice although source signals $L(j\omega)$ and $R(j\omega)$ may be accessible from the radio which produced them, the acoustic transfer functions $H_{AR}(j\omega)$ and $H_{AL}(j\omega)$ can only be estimated.

A simple approach to the estimation of the acoustic transfer function is to find long term ratio of microphone signal spectrum to each of the source stereo signals. Equations herein below describe this process for the right acoustic channel. Those skilled in the art will understand that a similar set of equations can be derived for the left acoustic channel. A basic transfer function H_{AR} for the right acoustic channel may be written as follows:

$$\hat{H}_{AR}(j\omega)=\frac{Y(j\omega)}{R(j\omega)}$$

Equation (1)

A spectrum of the signal $Y(j\omega)$ received at the microphone signal may be written as:

$$Y(j\omega)=H_{AR}(j\omega).R(j\omega)+H_{AL}(j\omega).L(j\omega)+S(j\omega)$$

Equation (2)

Substituting for $Y(j\omega)$ in equation (1) gives:

$$\hat{H}_{AR}(j\omega) = H_{AR}(j\omega) + H_{AL}(j\omega) \cdot \frac{L(j\omega)}{R(j\omega)} + \frac{S(j\omega)}{R(j\omega)}$$

Equation (3)

The following conclusions may be drawn from equation(3):

- In the case of a mono transmission being output through loudspeakers 101 and 102 whilst the user is saying a voice command, signals $L(j\omega)$ and $R(j\omega)$ are completely correlated with each other whilst being completed uncorrelated with $S(j\omega)$. In this case, individual left and right channel transfer functions cannot be uniquely determined, but a composite estimate which contains terms due to both left and right channels can be obtained. This is sufficient for practical cancellation of the mono ECAD signal output through the two loudspeakers received at the microphone.
- If $L(j\omega)$ and $R(j\omega)$ and $S(j\omega)$ are all uncorrelated, a correct estimate of the channel response will be obtained because second and third terms in equation (3) will normally have long term averages of O.
- If $L(j\omega)$ and $R(j\omega)$ are partially correlated, left and right acoustic channels cannot be unambiguously estimated. However, if $L(j\omega)$ and $R(j\omega)$ occupy different spectral regions or if corresponding time domain signals $l(t)$ and $r(t)$ have periods where one has low energy whilst the other has high energy, it may be still possible to make useful estimates of left and right channels for purposes of cancellation.

The frequency domain estimation of the right acoustic

channel response given by equation (3), and a corresponding equation for the left acoustic channel transfer function, $H_{AL}(j\omega)$, may be used to obtain an estimate of the magnitude of the wanted speech spectrum $S(j\omega)$. An estimate of the wanted speech magnitude spectrum may be obtained by subtracting the estimates of the left and right acoustic channels of the ECAD signals from the acoustic signal $Y(j\omega)$ received at the microphone:

$$\widehat{S^2}(\omega)=Y^2(\omega)-\widehat{H^2}_{AR}.R^2(\omega)-\widehat{H^2}_{AL}.L^2(\omega)$$

Equation (4)

An estimate of the acoustic channel power transfer function for the right acoustic channel, derived by squaring equation (3) may be as follows:

$$\widehat{H^2}_{AR}(\omega)=H^2_{AR}(\omega)+H^2_{AL}(\omega).\frac{L^2(\omega)}{R^2(\omega)}+\frac{S^2(\omega)}{R^2(\omega)}$$

Equation (5)

A corresponding estimate of the acoustic channel power transfer function for the left acoustic channel can also be derived by those skilled in the art.

Using an iterative approach, coupled with time and frequency dimension smoothing of the estimates of the channel response may be used to overcome problems caused by left and right signal correlation described herein above. Another problem which may need to be addressed arises because phase information in the channel response may be ignored, as the phase of the interferer is not normally required at the speech recognition system. As noted above,

cancellation for the purpose of speech recognition only requires an estimate of the magnitude of the speech spectrum because Mel Frequency Cepstral Co-efficient (MFCC) feature vector used by the speech recognition system in the preferred embodiment is based on magnitude spectra. The MFCC may be obtained by subjecting the speech spectrum in the frequency domain to a fast fourier transform in order to obtain its power in various frequency slots. The value of the power in the frequency domain is then passed through a log function and then a cosine transform to obtain the cepstrum in which the elements are orthogonal.

Normally, the phase characteristic encodes a frequency dependent delay spread associated with the acoustic transfer function. In a car typically the minimum delay is about 3ms. The delay spread may be compensated when making the channel estimate using equation (5). However, this compensation may be unnecessary if the spectral evaluation is done using a fast fourier transformer with block length much greater than the channel delay.

A practical form of the cancellation of non-stationary interferer signals such as those produced by ECAD may therefore be achieved using an algorithm 200 as illustrated by steps in Figure 2 of the accompanying drawings. In the preferred embodiment, the steps 201 to 205 are repeated once for each single frame (i.e a signal received at the microphone in a fixed period of time), however, initialisation steps 201 and 202 may only be performed for a first frame. At step 201, estimates of magnitudes of the left and right

channel transfer functions, $H_{AL}(J\omega)$ and $H_{AR}(J\omega)$ are initialised (set to zero):

$$\bar{H}^2_{AR}(\omega)=\bar{H}^2_{AL}(\omega)=0$$

At step 202, estimates of magnitude of left and right channel interference, C_L and C_R, are initialised:

$$C^2_{R,n-1}(\omega)=C^2_{L,n-1}(\omega)=0$$

At step 203, new estimates of magnitudes of the left and right interference signals at the microphone are calculated. This is achieved for the left microphone signal by subtracting the channel estimate of the magnitude of the right channel (calculated during the algorithm iteration for the immediately previous frame) from the microphone signal received at the current iteration (n). For the right interference channel, the magnitude estimate for the left channel derived during the previous iteration (n-1) is subtracted from the microphone signal:

$$C^2_{L,n}(\omega)=Y^2_n(\omega)-C^2_{R,n-1}(\omega)$$

(Equation 6)

$$C^2_{R,n}(\omega)=Y^2_n(\omega)-C^2_{L,n-1}(\omega)$$

(Equation 7)

At step 204, rough estimates of the left and right transfer functions, $H_{AL}(j\omega)$ and $H_{AR}(j\omega)$, are made. This is achieved for the left channel transfer function by dividing

the estimated left interference signal calculated at step 203 by the signal transmitted from the left stereo acoustic channel. For the right transfer function, the right channel interference signal estimate calculated at step 203 is divided by the signal transmitted from the right acoustic stereo channel:

$$\hat{H}^2_{AL,n}(\omega) = \frac{C^2_{L,n}(\omega)}{L^2_n(\omega)}$$

(Equation 8)

$$\hat{H}^2_{AR,n}(\omega) = \frac{C^2_{R,n}(\omega)}{R^2_n(\omega)}$$

(Equation 9)

Substituting equations (6) and (7) into the terms for the estimated interference signals in equations (8) and (9), respectively, gives expressions used to provide rough estimates of the left and right channel transfer functions:

$$\hat{H}^2{}_{ALn}(\omega) = \frac{\hat{Y}^2{}_n(\omega) - C^2{}_{R,n-1}(\omega)}{L^2{}_{n(\omega)}}$$

$$\hat{H}^2{}_{ARn}(\omega) = \frac{\hat{Y}^2{}_n(\omega) - C^2{}_{L,n-1}(\omega)}{R^2{}_{n(\omega)}}$$

At step 205 the rough estimates of the channel transfer functions obtained at step 204 may be smoothed, preferably both in the time and frequency domains. Time smoothing is preferably achieved with a first order recursive filter using a time constant of several hundred milliseconds. For example, time smoothing for the right channel may be as follows (a similar equation may also be obtained):

$$\overline{H}^2_{AR,n}=\beta\,.\,\overline{H}^2_{AR,n-1}+(1-\beta)\,.\,\hat{H}^2_{AR,n}$$

Frequency smoothing is preferably achieved using a Finite Impulse Response filter (represented by $f(\omega)$ in an equation herein below) with a triangular impulse response covering about 300 Hertz. Frequency smoothing for the right channel may be as follows (a similar expression for the left channel may also be obtained):

$$\tilde{H}_{AR,n}=f(\omega)*\overline{H}^2_{AR,n}(\omega)$$

The cancellation algorithm 200 described in steps 201 to 205 herein above may be refined by means of the four ways described herein below in order to attempt to deal with problems highlighted by equation (3) concerning correlation of left and right channel signals:

1. Updating of the recursive filter providing the smoothed channel estimate can be inhibited unless energy of one channel greatly exceeds energy of the other channel. This is preferably achieved by updating the left or right channel response only when it is assumed that only left or right channel, respectively, is active. Thus, a new right

acoustic channel transfer function would be estimated at step 204 if a ratio of the total energy of the signal transmitted from the right acoustic stereo channel by the total energy of the signal transmitted from the left stereo acoustic channel exceeds a predetermined threshold value, otherwise the estimate calculated for the transfer function during the previous frame iteration is used. A corresponding estimation would also be performed for the left transfer function.

Using E_L to represent the total energy in the n_{th} frame of the left stereo acoustic channel and E_R represent the total energy in the n_{th} frame of the right stereo acoustic channel. Thus, the channel response estimation algorithm for the right channel is:

$$\hat{H}_{AR,n} = \frac{Y(j\omega)}{R(j\omega)} \; if \; \frac{E_R}{E_L} \geq Threshold$$

otherwise use previous estimate $(\hat{H}_{AR,n-1})$ if $E_R/E_L <$ Threshold.

The channel response estimation algorithm for the left channel is:

$$\hat{H}_{AL,n} = \frac{Y(j\omega)}{L(j\omega)} \; if \; \frac{E_L}{E_R} \geq Threshold,$$

otherwise use previous estimate $(\hat{H}_{AL,n-1})$ if $E_R/E_L <$ Threshold.

Normally, when considering the right channel, when the threshold is exceeded, Y(jω) should consist mainly of terms due to the right channel and the wanted speech signal. Y(jω) should contain very little energy due to the left channel if the threshold is set at high value. The reverse normally holds when considering the left channel. Time and domain smoothing substantially as described at step 205 would also be used.

2. Updating of recursively smoothed channel estimate at particular frequencies can be inhibited unless energy at that frequency in one channel greatly exceeds the energy at that frequency in the other channel. This may be achieved by estimating new values for the left and/or right acoustic channel transfer functions when a ratio of the total energies of the left and right stereo acoustic signals exceeds a given threshold at individual frequency components in the spectrum. Preferably, the threshold may apply to frequencies comprising a harmonic number in the Discreet Fourier Transforms of the signals.

Using a similar terminology to that in 1. herein above, the channel response estimation algorithm for the right channel is:

$$\hat{H}_{AR,n}(k) = \frac{Y(k)}{R(k)} \; if \frac{E(k)_R}{E(k)_L} \geq Threshold$$

Otherwise use estimate at previous iteration ($\hat{H}_{AR,n-1}$) if $E(k)_R/E(k)_L <$ Threshold.

The channel response estimation algorithm for the left channel is:

$$\hat{H}_{AL,n}(k) = \frac{Y(k)}{L(k)} \; if \; \frac{E(k)_L}{E(k)_R} \geq Threshold$$

otherwise use the estimate calculated at the previous iteration ($\hat{H}_{AL,n-1}$) if $E(k)_L/E(k)_R$ < Threshold.

In this definition, the index k refers to the harmonic number in the DFTs of the signals. For example, $E(k)_R$ is the energy of the kth harmonic in the DFT of the right stereo source signal. This algorithm should ensure that the acoustic channel responses are only updated at those frequencies and at those time at which the signal at the microphone consists mainly of either left or right channel.

3. Evaluate coherence function between the left and right channel signals and use inverse magnitude of the coherence at each frequency as a weighting on the amount by which estimates of the channel responses are updated at that frequency. The coherence function provides a measure of correlation over a period of time of phases of two different signals measured at a particular frequency. The coherence function may be used in various ways, normally based on the idea that the update of the acoustic channel responsible will be decreased if the left and right stereo channels are phase-correlated at a particular frequency. If the coherence approaches unity, the signals are correlated, but only

at the specified frequency. Thus, the channel response estimates for the right channel may be derived from the following algorithm (a corresponding method for the transfer function for the left channel may also be derived):

$$\hat{H}^{2}_{AR}(k) = \frac{Y(k)}{R(k)} \,.\, (1-|\eta(k)|)$$

where $\eta(k)$ is the coherence of the left and right stereo source signal at frequency index k.

$$\eta(k) = \left\langle \frac{L(k) \,.\, R^{*}(k)}{|L(k)| \,.\, |R(k)|} \right\rangle$$

where the expectation is over time.

4. Extract those components of the left and right ECAD source signals which are uncorrelated (orthogonal) and use them to make estimates of the left and right channel responses. In this approach, a common component C(k) in the left and right ECAD sources is removed by adaptive filtering to yield an orthogonal pair of signals, L''(k) and R''(k):

$R(k)=R''(k)+H_{CR}(k).C(k)$

$L(k)=L''(k)+H_{CL}(K).c(k)$

wherein $H_{CL}(k)$ is the transfer function between the common (left and right stereo signals combined, which may be fixed in a recording studio) ECAD signal source and the left ECAD signal source and $H_{CR}(k)$ is the transfer function between the common ECAD source, signal and the right ECAD source.

The orthogonalised signals are used to make the acoustic channel response estimates. For the right stereo channel transfer function the following expression may be used (a corresponding expression for the left stereo channel transfer function may also be obtained):

$$\hat{H}_{AR}(k)=\frac{Y(k)}{R''(k)}=\frac{H_{AR}(K).(R''(k)+H_{CR}(k).C(k))}{R''(k)}+\frac{H_{AL}(k).(L''(k)+H_{CL}(k).C(k))}{R''(k)}+\frac{S(k)}{R''(k)}$$

Most of the terms are long term uncorrelated so we get:

$$\hat{H}_{AR}(k)=\frac{Y(k)}{R''(k)}=\frac{H_{AR}(k).R''(k)}{R''(k)}=H_{AR}(k)$$

the true acoustic channel response.

Thus, the right stereo acoustic channel function, $\hat{H}_{AR}(k)$, may be obtained by dividing the signal received at the microphone by R''(k).

Figure 3 of the accompanying drawings illustrates schematically an example of components which may be used to form L''(jω) and R''(jω). The components include two adaptive filters, 303 and 304, either implemented in the frequency domain, or preferably, the time domain. The coefficients of each FIR adaptive filter are adjusted using LMS or similar, to minimise the total energy in r''(n) and l''(n), respectively, i.e. operate filters in standard system identification mode as in echo cancelling etc.

The right stereo ECAD signal r(n) 301 is fed into

adaptive filter 303 and a combiner 305. The left stereo ECAD signal l(n) 302 is fed into adaptive filter 304 and a combiner 306. The output of adaptive filter 303 is also fed into combiner 306. The output of adaptive filter 304 is also fed into combiner 305. The output of combiner 305 may be fed back via an adaption control path into adaptive filter 304. The output of mixer 306 may be fed back into adaptive filter 303 via an adaption control path. The output of combiner 305 comprises the orthogonal right stereo signal r''(n) 307. The output of combiner 306 comprises the left stereo orthogonal signal l''(n) 308.

Figure 4 of the accompanying drawings illustrates a block diagram representing a specific embodiment of the present invention. Processing components of Fig. 4 may be electronic processors fitted integrally to the in-car device where the speech recognition system is located or, alternatively, may be a stand alone electronic device intended to receive acoustic signals, cancel non-stationary interfering signals and output a filtered acoustic signal to be received by the speech recognition system's microphone.

ECAD sound source 401 (such as the signals output loudspeakers 101 and 102 of Figure 1) may be received directly by a spectral analysis process 404 so that the signal as produced by the ECAD prior to transmission through the in-car acoustic channel 403 may be analysed. The ECAD signal is also received by a spectral analysis process 405 after transmission through acoustic channel 403 so that the signal 401 is in effect simultaneously spectrally analysed

before and after transmission through the acoustic channel 403. The spectral analysis of processes 404 and 405 is preferably carried out at a 16 ms frame rate using a 256 point Fast Fourier Transformer. If user speech 402 (corresponding to wanted speech signal $S(j\omega)$ 104 of Figure 1) is also present then this acoustic signal too will also be transmitted through the acoustic channel 403 and received by spectral analysis process 405.

The output of spectral analysis processes 404 and 405 are used as inputs to acoustic channel model estimation process 406 which preferably functions in accordance with algorithm 200 described herein above. Acoustic channel model estimation process 406 produces an acoustic channel model 407 which may be used as an input to a spectral subtraction process 408 which also receives the acoustic signal transferred through channel 403.

When the speech recognition system is required, the acoustic channel model 407 is frozen for duration of the speech recognition process. The acoustic channel model 407 is then used to recover the speech signal from the microphone signal by subtracting the estimated spectrum of the ECAD interfering signals contained in the model 407 from the acoustic signals received at the microphone. The spectrally subtracted signal representing the recovered wanted speech 409 is then passed to a pattern matcher process 410 (part of the speech recognition system) which may use recognition feature sets such as Hidden Markov of models 311 in order to match the recovered speech signal 409 to a command which is

recognised by the system. The pattern matcher 409 may then pass on an output signal to trace back and decision process 412 in order that the user's speech command be carried out by the device.

Since the spectral subtraction algorithm is frame rather than sample based, its computational complexity is low. The algorithm's main computation is required for the Fast Fourier Transform, which requires order NlogN computations per frame for each channel. This is typically only about 250k computations per second, which is significantly lower than the order 3N computations per sample required by the simplest form of known adaptive filter technique. For an echo tail length of 32 microseconds, 256 samples, this equates to more than 18 million operations per second.

Figures 5 to 8 of the accompanying diagrams illustrate microphone signal traces before and after the non-stationary interferer signal cancellation for different types of music output by the ECAD at different signal to interference ratios. In order to allow for comparison between an uncancelled signal passed through the acoustic channel and the cancelled signal, test data was constructed by recording speech and interferer signals separately in the same car environment and then adding the two signals. In the examples shown in figures 5 to 8, the interfering music is a stereo signal.

Figures 5A to 5D of the accompanying drawings illustrate microphone traces with and without cancellation in a case where the ECAD outputs pop music at 0dB signal to

interference ratio. In Fig. 5A a signal received at the microphone prior to cancellation is illustrated. In this case, peak segmental speech and interferer levels are the same. This is a highly pessimistic way of estimating signal-to-noise ratio as amplitude variability of speech signal is higher than that of the ECAD music signal output which exceeds the speech for a considerable part of the example. Fig. 5B illustrates a signal resulting from an inverse transformation on the signal of Fig. 5A after spectral subtraction. The interfering signal as shown in Fig. 5B has clearly been reduced. Fig. 5C illustrates a signal representing normalised squared cepstral distances for application of the cancellation algorithm. Fig. 5D illustrates a signal trace for the normalised squared cepstral distances of Fig. 5C after spectral subtraction. Comparing the traces illustrated in Fig. 5C and 5D, it can be seen that the recovered speech cepstral are less distorted than with the interferer.

Figures 6A to 6D of the accompanying drawings illustrate microphone traces with and without cancellation in a case where the ECAD outputs pop music at 10 decibel signal to interference ratio. In Fig. 6A a signal received at the microphone prior to cancellation is illustrated. Fig. 6B. illustrates a signal resulting from an inverse transformation on the signal of 6A after spectral subtraction. The interfering signal shown in Fig. 6B has clearly been reduced. Fig. 6C illustrates a signal representing normalised squared cepstral distances for application of the

cancellation algorithm. Fig. 6D illustrates a signal trace for the normalised square cepstral distances of Fig. 6C after spectral subtraction.

Figures 7A to 7D of the accompanying drawings illustrate microphone traces with and without cancellation in a case where the ECAD outputs opera music at 0 decibel signal to interference ratio. In Fig. 7A a signal received at the microphone prior to cancellation is illustrated. Fig. 7B. illustrates a signal resulting from an inverse transformation on the signal of 7A after spectral subtraction. The interfering signal shown in Fig. 7B has clearly been reduced. Fig. 7C illustrates a signal representing normalised squared cepstral distances for application of the cancellation algorithm. Fig. 7D illustrates a signal trace for the normalised square cepstral distances of Fig. 7C after spectral subtraction.

Figures 8A to 8D of the accompanying drawings illustrate microphone traces with and without cancellation in a case where the ECAD outputs opera music at 10 decibel signal to interference ratio. In Fig. 8A a signal received at the microphone prior to cancellation is illustrated. Fig. 8B. illustrates a signal resulting from an inverse transformation on the signal of 8A after spectral subtraction. The interfering signal shown in Fig. 8B has clearly been reduced. Fig. 8C. illustrates a signal representing normalised squared cepstral distances for application of the cancellation algorithm. Fig. 8D illustrates a signal trace for the normalised square cepstral distances of Fig. 8C

after spectral subtraction.

Claims

1. Apparatus for cancellation of one or more non-stationary interfering signals for speech recognition, said apparatus comprising:

means for receiving an acoustic signal;

means for generating an estimated value of a magnitude spectrum of said non-stationary interfering signals; and

means for subtracting said estimated value from said received acoustic signal to produce a representation of a wanted speech magnitude spectrum.

2. Apparatus according to claim 1, wherein said means for generating estimated value includes processing means configured to estimate a transfer function for an acoustic channel between each source of said non-stationary interfering signals and said means for receiving an acoustic signal.

3. Apparatus according to claim 2, wherein said processing means is configured to estimate transfer functions for said non-stationary interfering signals produced by left and right stereo channel transmissions.

4. Apparatus according to Claim 2 or Claim 3, wherein said estimation of said transfer functions is achieved by said processing means executing an iterative algorithm on a frame-by-frame basis, the frames being constituted by said acoustic signals received during successive time periods.

5. Apparatus according to Claim 4 when dependent upon Claim 3, wherein said processing means is configured to estimate respective magnitudes of said left and right channel interference signals,

said magnitude of left channel interference signal is estimated by subtracting said right channel interference signal magnitude estimated during previous said iteration from said acoustic signal received at current said iteration; and

said magnitude of right channel interference signal is estimated by subtracting said left channel interference signal magnitude estimated during previous said iteration from said acoustic signal received at current said iteration.

6. Apparatus according to Claim 5, wherein said transfer function estimate for said right stereo acoustic channel is determined by dividing said right channel interference magnitude estimate by said interfering signal transmitted from said right acoustic stereo channel; and

said transfer function estimate for said left stereo acoustic channel is determined by dividing said left channel interference magnitude estimate by said interfering signal transmitted from said left acoustic stereo channel.

7. Apparatus according to Claim 6, wherein said right acoustic channel transfer function estimation is performed for a said iteration only if a ratio of total energy of said right acoustic stereo channel interfering signal over total energy of said left acoustic stereo channel interfering signal exceeds a predetermined threshold value; and

said left acoustic channel transfer function estimation is performed for a said iteration only if a ratio of total energy of said left acoustic stereo channel interfering

signal over total energy of said right acoustic stereo channel interfering signal exceeds a predetermined threshold value.

8. Apparatus according to Claim 7, wherein said ratio and threshold comparisons are applied to individual frequency components in spectra of said signals.

9. Apparatus according to Claim 8, wherein said left and right stereo acoustic channel transfer functions are multiplied by $(1-|\eta(k)|)$ where $\eta(k)$ is coherence of said left and right interfering signals at a frequency index k.

10. Apparatus according to Claim 4, wherein said transfer function estimate for said right stereo acoustic channel is obtained using an expression:

$$\hat{H}_{AR}(k) = \frac{Y(k)}{R''(k)} = \frac{H_{AR}(k) \, . R''(k)}{R''(k)} = H_{AR}(k)$$

and said transfer functions estimate for said left stereo acoustic channel is obtained using an expression:

$$\hat{H}_{AL}(k) = \frac{Y(k)}{L''(k)} = \frac{H_{AL}(k) \, . L''(k)}{L''(k)} = H_{AL}(k)$$

wherein $R''(k)=H_{CR}(k).C(k)$, with $C(k)$ being a common component of said left and right stereo channel signals and $H_{CR}(k)$ is a transfer function between common said left and right stereo channel transmissions, and said right stereo channel and $L''(k)=L(k)-H_{CL}(k).C(k)$, where $H_{CL}(k)$ is a transfer function between common said left and right stereo channel transmissions and said left stereo channel signal.

11. Apparatus according to any one of claims 2 to 10, wherein said processing means further comprises means for smoothing said estimated transfer functions in time domain.

12. Apparatus according to claim 11, wherein said means for smoothing in time domain comprises a first order recursive filter.

13. Apparatus according to any one of claims 2 to 12, wherein said processing means further comprises means for smoothing said estimated transfer functions in frequency domain.

14. Apparatus according to Claim 13, wherein said means for smoothing in frequency domain comprises a Finite Impulse Response filter.

15. Apparatus according to any one of claims 2 to 14, wherein said processing means includes means for performing a Fourier Transform.

16. Apparatus according to any of the preceding claims, wherein said non-stationary interfering signals are produced by an electronic acoustic device operating in a vehicle.

17. Apparatus according to any one of the preceding claims, wherein said means for receiving an acoustic signal comprises a microphone.

18. A method of cancellation of one or more non-stationary interfering signals for speech recognition, said method comprising steps of:

receiving an acoustic signal;

generating an estimated value for a magnitude spectrum of said non-stationary interfering signal; and

subtracting said estimated value from said received acoustic signal to produce a representation of a wanted speech magnitude spectrum.

19. Method according to Claim 18, wherein said step of generating an estimated value comprises estimating a transfer function for an acoustic channel between each source of said non-stationary interfering signals and said means for receiving an acoustic signal.

20. Method according to Claim 19, wherein said transfer functions are estimated for non-stationary interfering signals produced by left and right stereo channel transmissions.

21. Method according to any one of Claims 18 to 20, wherein said steps are executed iteratively on a frame-by-frame basis, the frames being constituted by said acoustic signals received during successive time periods.

22. Method according to Claim 21, when dependent upon Claim 20, wherein said step of estimating a transfer function includes:

estimating a magnitude of said left channel interference signal by subtracting said right channel interference signal magnitude estimated during previous said iteration from said acoustic signal received at current said iteration; and

estimating magnitude of said right channel interference signal by subtracting said left channel interference signal magnitude estimated during previous said iteration from said acoustic signal received at current said iteration.

23. Method according to Claim 22, further comprising steps of:

dividing said right channel interference magnitude estimate by said interfering signal transmitted from said right acoustic stereo channel; and

dividing said left channel interference magnitude estimated by said interfering signal transmitted from said left acoustic stereo channel.

24. Method according to Claim 23, wherein said step of estimating right acoustic channel transfer function is performed for a said iteration only if a ratio of total energy of said right acoustic stereo channel interfering signal over total energy of said left acoustic stereo channel interfering signal exceeds a predetermined threshold value; and

said step of estimating left acoustic channel transfer function estimate is performed for a said iteration only if a ratio of total energy of said left acoustic stereo channel interfering signal over total energy of said right acoustic stereo channel interfering signal exceeds a predetermined threshold value.

25. Method according to Claim 24, wherein said ratio and threshold comparisons are applied to individual frequency components in spectra of said signals.

26. Method according to Claim 25, wherein said left and right stereo acoustic channel transfer functions are multiplied by $(1-|\eta(k)|)$ where $\eta(k)$ is coherence of said left and right interfering signals at a frequency index k.

27. Method according to Claim 21, wherein said transfer function estimate for said right stereo acoustic channel is obtained using an expression:

$$\hat{H}_{AR}(k)=\frac{Y(k)}{R''(k)}=\frac{H_{AR}(k)\,.\,R''(k)}{R''(k)}=H_{AR}(k)$$

and said transfer functions estimate for said left stereo acoustic channel is obtained using an expression:

$$\hat{H}_{AL}(k)=\frac{Y(k)}{L''(k)}=\frac{H_{AL}(k)\,.\,L''(k)}{L''(k)}=H_{AL}(k)$$

wherein $R''(k)=H_{CR}(k).C(k)$, with $C(k)$ being a common component of said left and right stereo channel signals and $H_{CR}(k)$ is a transfer function between common said left and right stereo channel transmissions, and said right stereo channel and $L''(k)=L(k)-H_{CL}(k).C(k)$, where $H_{CL}(k)$ is a transfer function between common said left and right stereo channel transmissions and said left stereo channel signal.

28. Method according to any one of Claims 18 to 27, further comprising a step of smoothing said estimated transfer functions in time domain.

29. Method according to any one of Claims 18 to 28, further comprising a step of smoothing said estimated transfer functions in frequency domain.

30. A speech recognition system including apparatus according to any one of claims 1 to 17.

31. An electronic acoustic device including apparatus according to any one of claims 1 to 17.

Abstract

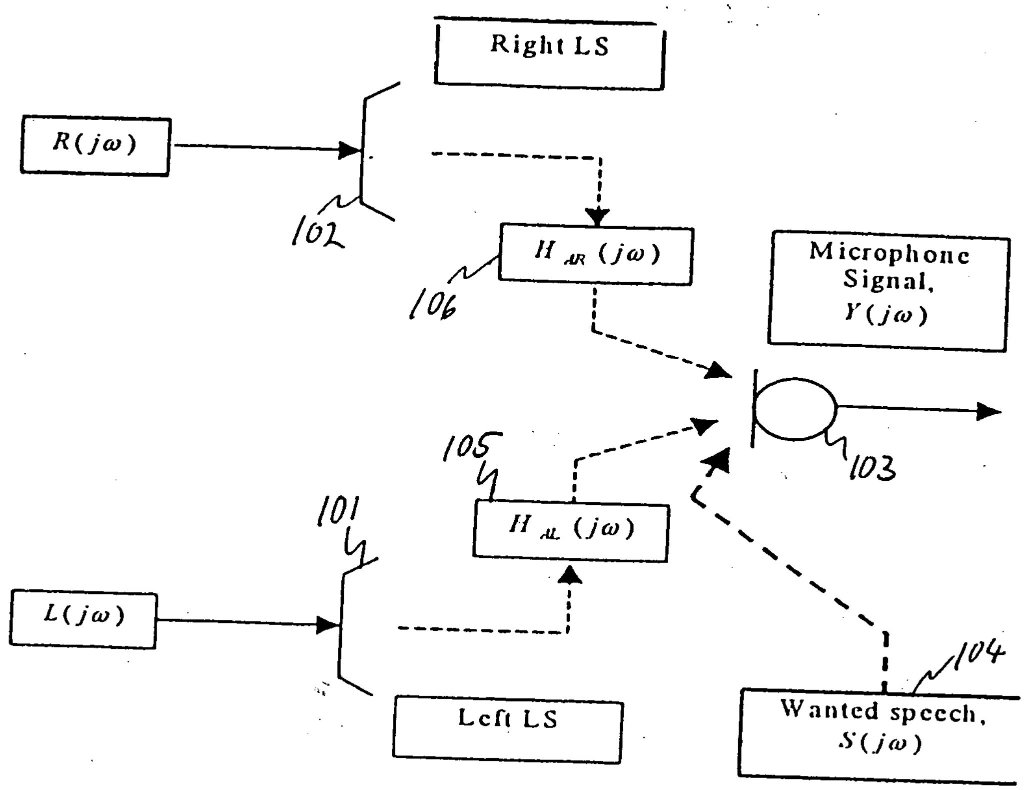
Cancellation of non-stationary interferer signals

for speech recognition

System for cancellation of non-stationary interfering signals, particularly for use for mitigating effects of such interferers produced by in-car entertainment (ECAD) devices for speech recognition applications. The system spectrally analyses signals output by the ECAD before and after they are passed through an in-car acoustic channel. A model of the acoustic channel is built by the system's algorithm. For speech recognition the model is spectrally subtracted from a signal received at a microphone in order to recover a wanted speech signal. The acoustic channel model is built by estimating frequency domain acoustic transfer functions between each loudspeaker used by the ECAD and the microphone.

Figure 1.

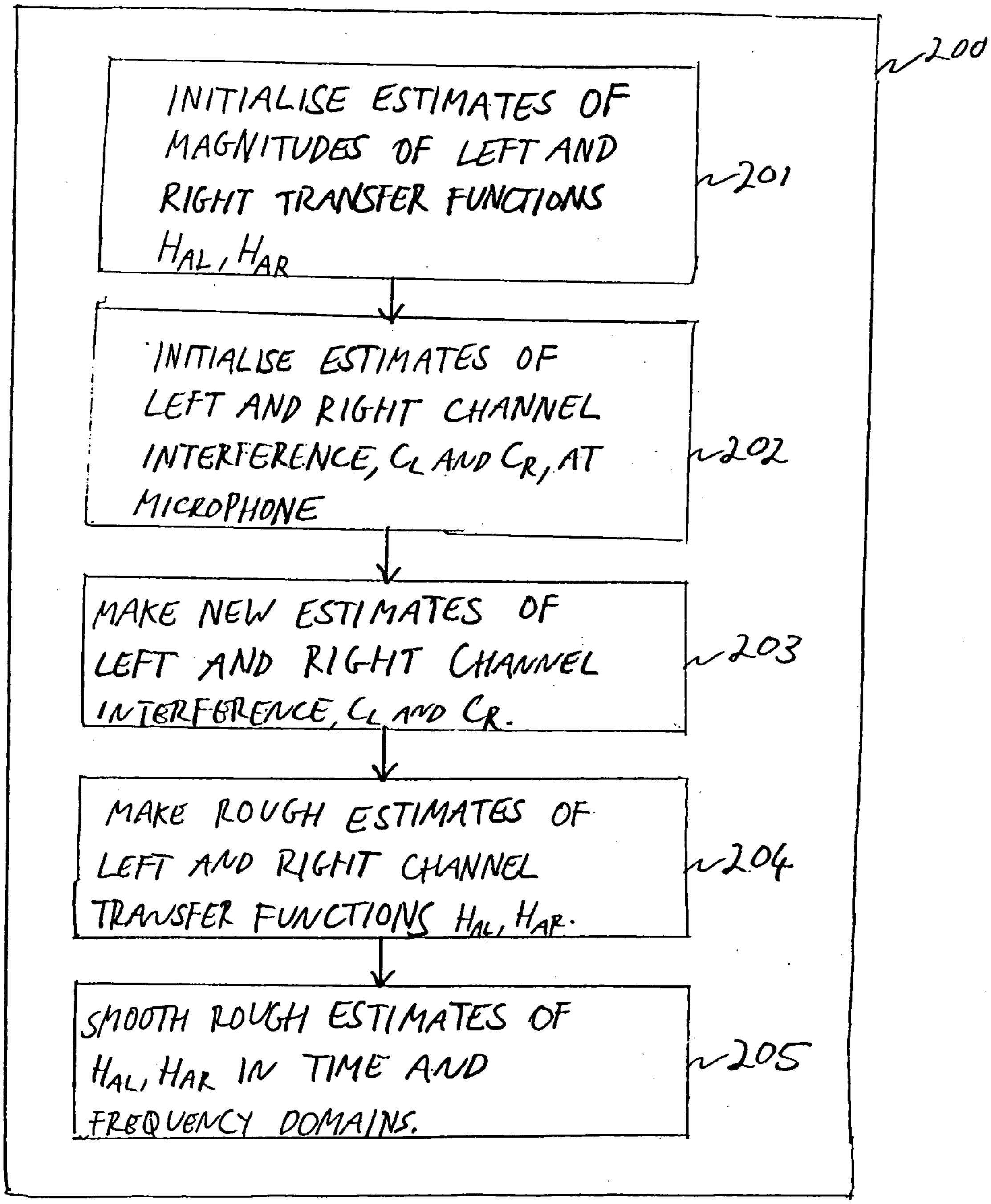
1/5



Right LS
R(jω)
102
H AR (jω)
106
Microphone Signal, Y(jω)
103
105
H AL (jω)
101
L(jω)
Left LS
104
Wanted speech, S(jω)

FIG. 1

THIS PAGE BLANK (USPTO)



200
INITIALISE ESTIMATES OF MAGNITUDES OF LEFT AND RIGHT TRANSFER FUNCTIONS H_{AL}, H_{AR}
201
INITIALISE ESTIMATES OF LEFT AND RIGHT CHANNEL INTERFERENCE, C_L AND C_R, AT MICROPHONE
202
MAKE NEW ESTIMATES OF LEFT AND RIGHT CHANNEL INTERFERENCE, C_L AND C_R.
203
MAKE ROUGH ESTIMATES OF LEFT AND RIGHT CHANNEL TRANSFER FUNCTIONS H_{AL}, H_{AR}.
204
SMOOTH ROUGH ESTIMATES OF H_{AL}, H_{AR} IN TIME AND FREQUENCY DOMAINS.
205

FIG. 2

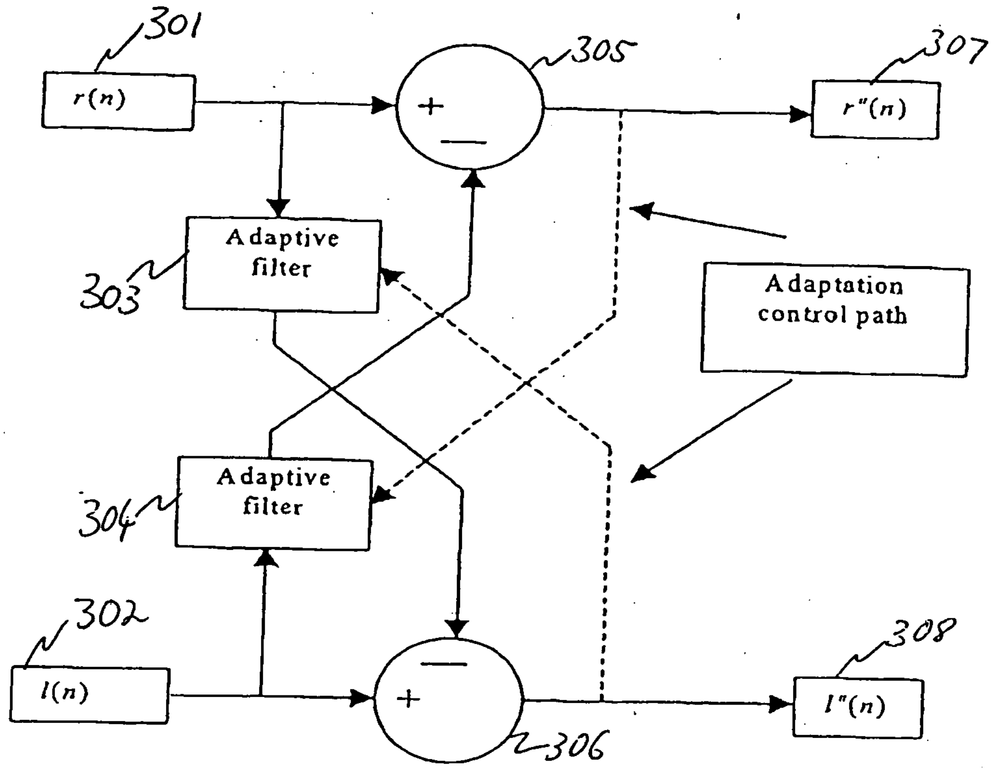

301
r(n)
+
305
307
r"(n)
Adaptive filter
303
Adaptation control path
Adaptive filter
304
302
l(n)
+
306
308
l"(n)

FIG.3

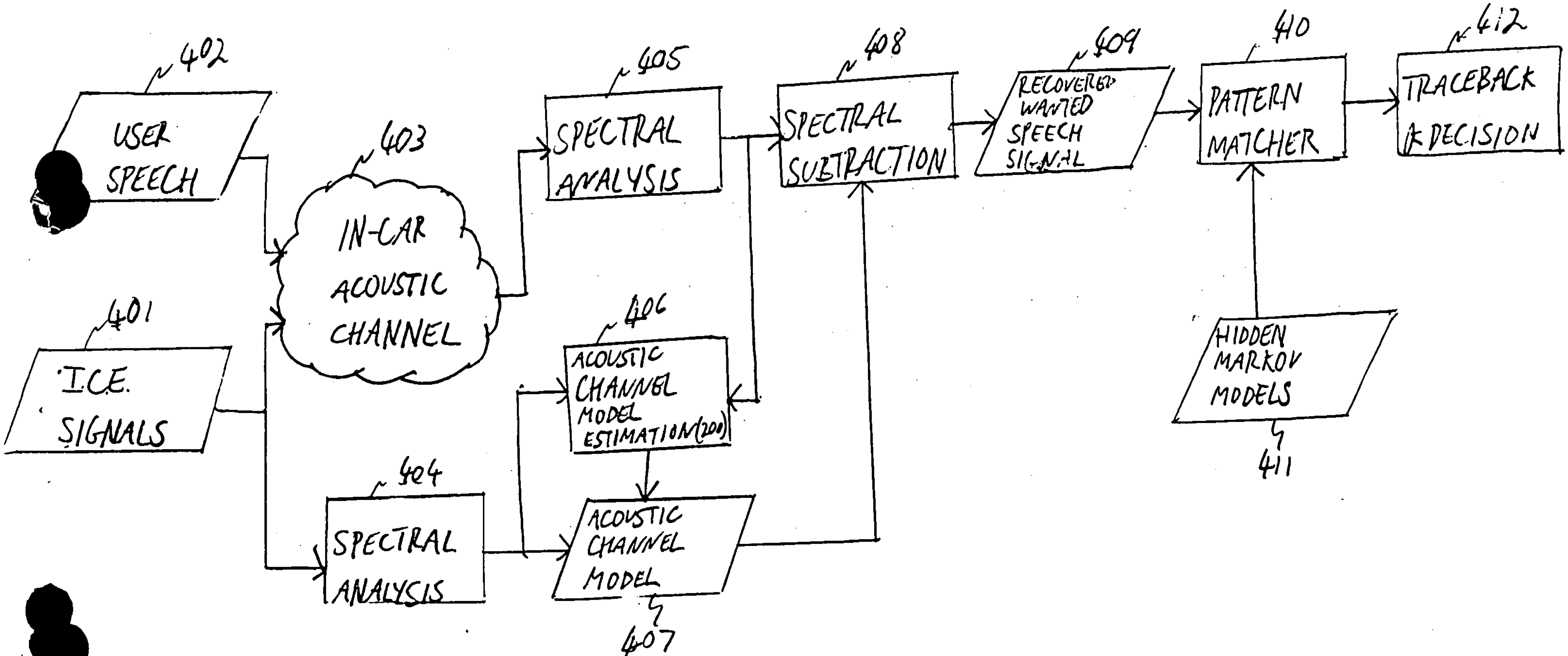

402
USER SPEECH
401
I.C.E. SIGNALS
403
IN-CAR ACOUSTIC CHANNEL
405
SPECTRAL ANALYSIS
404
SPECTRAL ANALYSIS
406
ACOUSTIC CHANNEL MODEL ESTIMATION(200)
ACOUSTIC CHANNEL MODEL
407
408
SPECTRAL SUBTRACTION
409
RECOVERED WANTED SPEECH SIGNAL
410
PATTERN MATCHER
HIDDEN MARKOV MODELS
411
412
TRACEBACK & DECISION

FIG.4

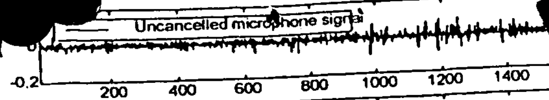
FIG 5A


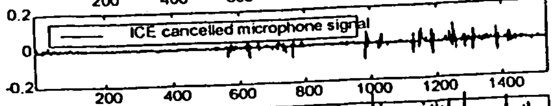
FIG 5B


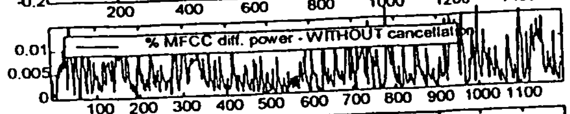
FIG. 5C


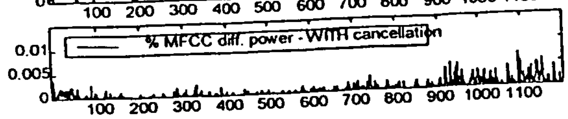
FIG 5D


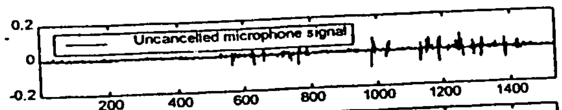
FIG. 6A


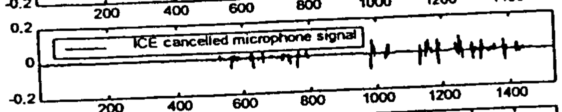
FIG. 6B


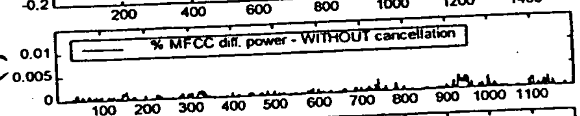
FIG. 6C


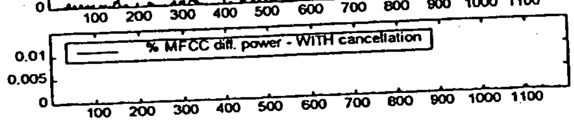
FIG 6D


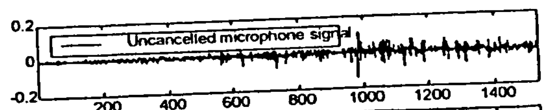
FIG. 7A


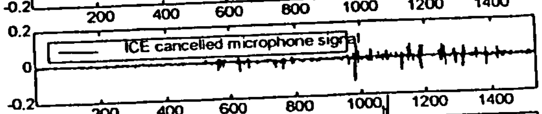
FIG. 7B


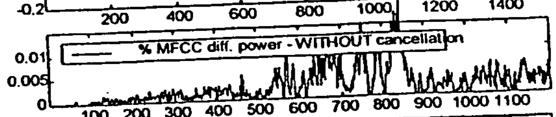
FIG. 7C


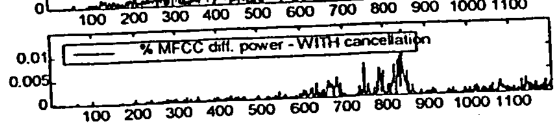
FIG. 7D


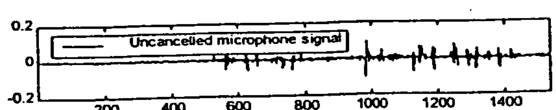
FIG. 8A


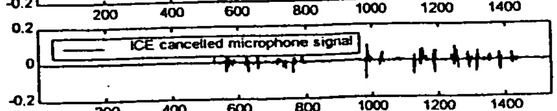
FIG. 8B


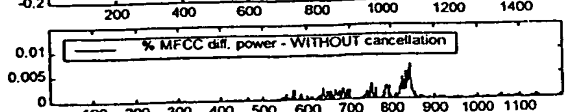
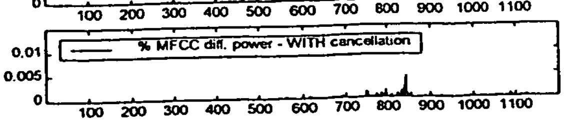
FIG. 8C


FIG. 8D


PCT/GB 00/01715
24/5/2000

WYNNE-JONES LAINE & JAMES